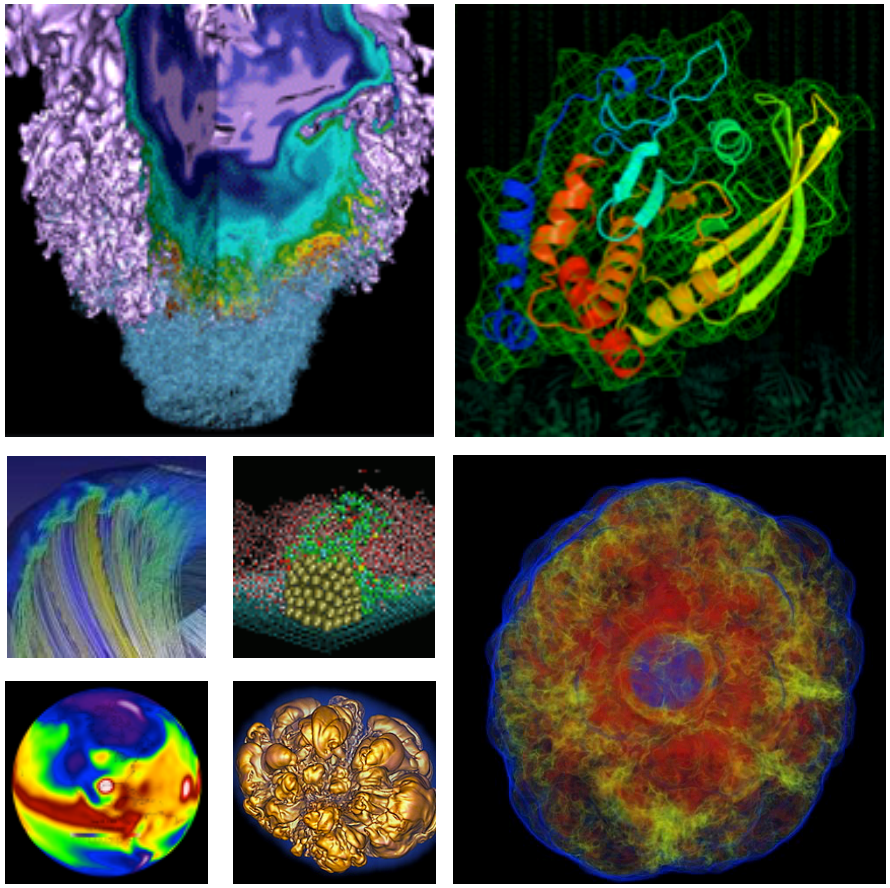
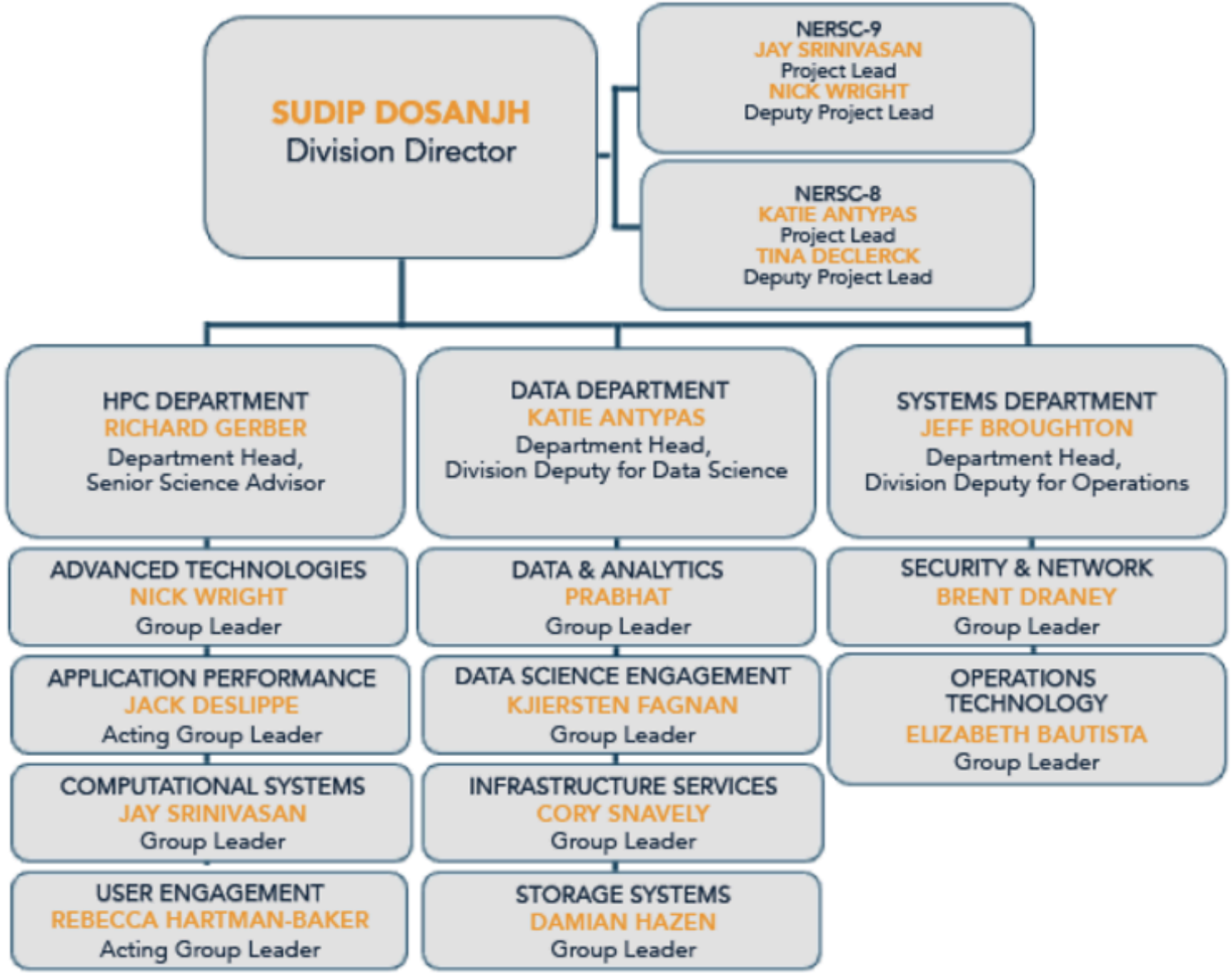
Data and Analytics Services Overview



Prabhat
DAS Group Lead

New User Training, 2/24/2017

NERSC Org Chart



DAS Team

Debbie Bard
Cosmology, Particle Physics
Workflows, Analytics

Wahid Bhimji
Particle Physics
Management, Transfer

Shane Canon
Systems Biology, Genomics
Workflows

Shreyas Cholia
Access, Management

Lisa Gerhardt
Particle Physics
Management, Workflows

Annette Greiner
BioMedical Sciences
Access, Visualization

Burlen Loring
Visualization

Jialin Liu
Management

Jeff Porter
Nuclear Physicsr
Transfer

Evan Racah
Analytics

Rollin Thomas
Astrophysics, Cosmology
Analytics, Access

Quincey Koziol
Management

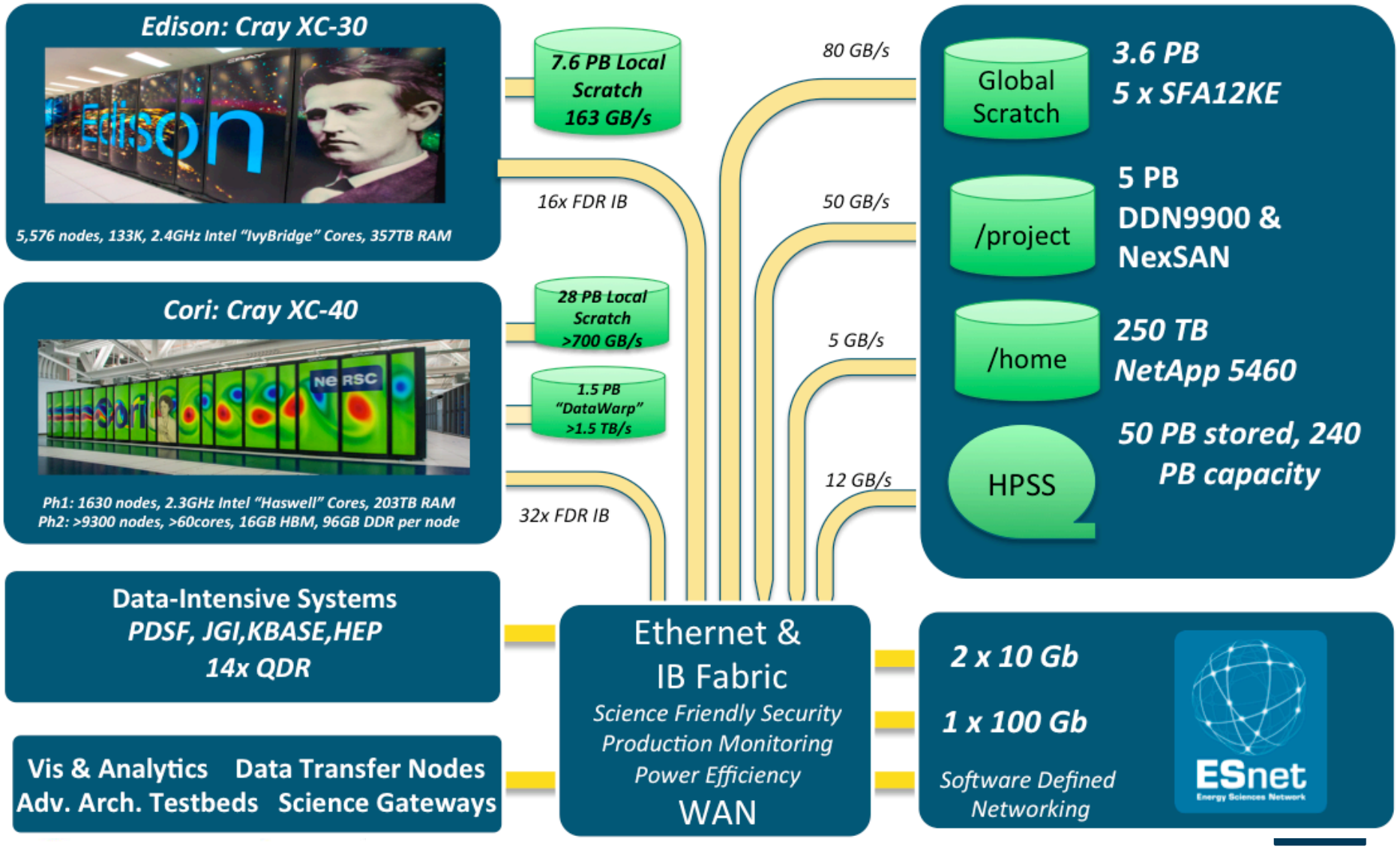
Prabhat
Climate, Neuroscience
Analytics, Management

DAS Goal:
"Enable Data-Intensive Science at Scale"

Internal Goals

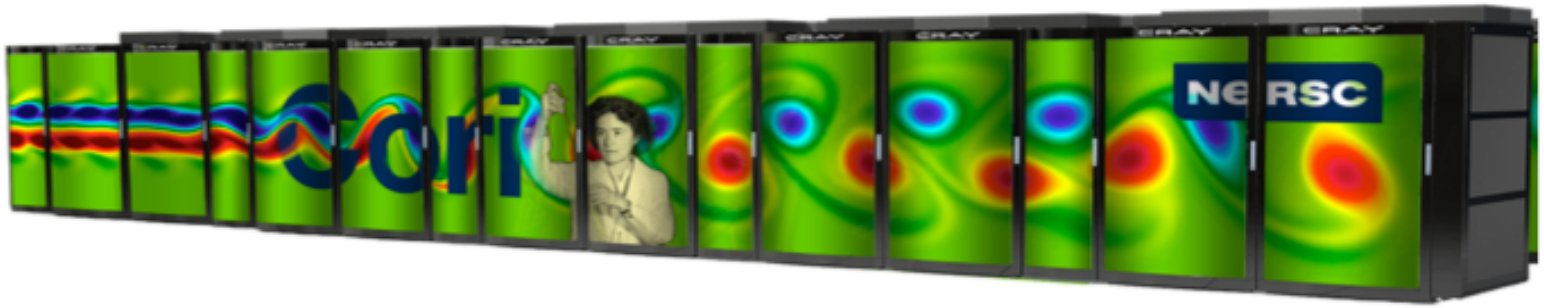
- Provide world-class, production quality software services for all major Data capabilities:
  - Analytics, Management, Workflows, Transfer, Access, Visualization
- Pioneer evaluation, research and deployment of Big Data technologies
  - Focusing on productivity and performance
- Engage with stakeholders to enable scientific discovery in a data-driven world
  - Users, Computing Sciences Staff, Vendors, Researchers

NERSC - 2017


Edison: Cray XC-30
5,576 nodes, 133K, 2.4GHz Intel "IvyBridge" Cores, 357TB RAM
Cori: Cray XC-40
Ph1: 1630 nodes, 2.3GHz Intel "Haswell" Cores, 203TB RAM
Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node
7.6 PB Local Scratch 163 GB/s
16x FDR IB
28 PB Local Scratch >700 GB/s
1.5 PB "DataWarp" >1.5 TB/s
32x FDR IB
80 GB/s
50 GB/s
5 GB/s
12 GB/s
Global Scratch
3.6 PB 5 x SFA12KE
/project
5 PB DDN9900 & NexSAN
/home
250 TB NetApp 5460
HPSS
50 PB stored, 240 PB capacity
Data-Intensive Systems
PDSF, JGI,KBASE,HEP
14x QDR
Vis & Analytics Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways
Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN
2 x 10 Gb
1 x 100 Gb
Software Defined Networking
ESnet
Energy Sciences Network

Cori Data Features

- **NVRAM Flash Burst Buffer as I/O accelerator**
- **High bandwidth external connectivity to experimental facilities from compute nodes (Software Defined Networking)**
- **More login nodes for managing advanced workflows**
- **Data-friendly queues**
  - Shared, Interactive, Real-time, High-mem, Transfer
- **Virtualization capabilities with Shifter**



Data Software and Services

Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, php, python django, newt
Workflows	FireWorks, swift
Data Management	HDF, netCDF, ROOT Data Analysis Framework, PostgreSQL, SciDB, mongoDB, MySQL
Data Analytics	python, jupyter, IPython, theano, Spark, R, TensorFlow, Caffe, MATLAB, Julia, Wolfram Mathematica 10
Data Visualization	VisIt, ParaView

How to get help

- **Documentation:**
  - www.nersc.gov
  - http://www.nersc.gov/users/data-analytics/
- **File a ticket**
  - send e-mail to consult@nersc.gov
- **NUG monthly telecons/webinars, Data Day**
- **In-person engagement**

Training Schedule

Time (PDT)	Topic	Presenter
9:00 am	Data and Analytic Services Overview	Prabhat
9:20 am	Python and ipython	Rollin Thomas
9:40 am	Machine Learning	Evan Racah
10:00 am	Spark	Lisa Gerhardt
10:20 am	Break	
10:40 noon	Databases and Data Formats	Wahid Bhimji, Quincey Kozoil, Jialin Liu
11:10 pm	Burst Buffer	Wahid Bhimji
11:30 pm	Lunch	
12:50 pm	Intro to ESNet	Eli Dart
1:20 pm	Moving and sharing data	Shreyas Cholia
1:40 pm	Shifter	Shane Canon
2:00 pm	Workflow Tools	Shane Canon
2:20 pm	Break	
2:40 pm	Science Gateways	Annette Greiner
3:00 pm	Visualization	Annette Greiner
3:20 pm	Q & A Session	
3:30 pm	End	

Asks…

- **Please engage with the DAS team**
  - Ask questions
  - Provide critical feedback on choice of tools
  - Let us know if we are missing capabilities
- **Please let us know of your use cases**
  - Productivity
  - Performance
    - Portfolio should work at 1TB / 10,000 core level
    - Scale to 100TB / 100,000 cores
- **Have Fun!**

National Energy Research Scientific Computing Center